Statistical Challenges in 21st Century Cosmology

IfA Edinburgh

Photometric redshifts by machine learning: Don’t underestimate your scatter & bias

Chieh-An Lin (Linc)

May 24th, 2018

Valencia, Spain

Galaxy spectrum redshifts

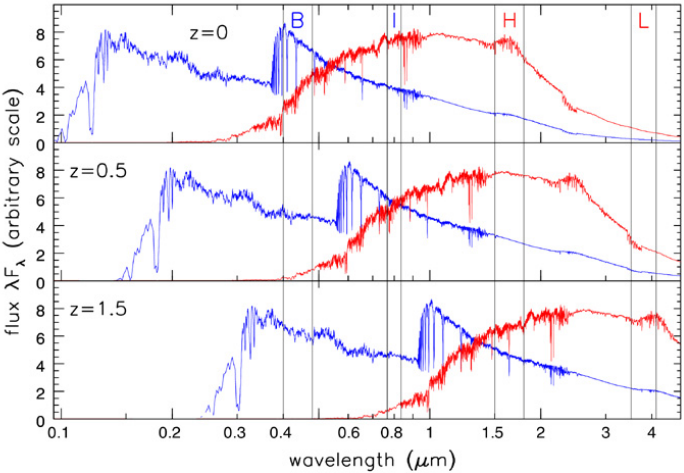
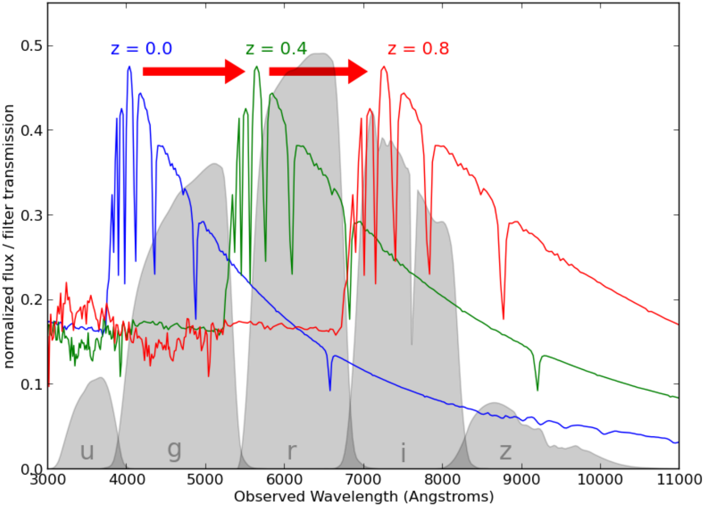


Fig 8.12 (S. Charlot) 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

Redshifted spectrum in photometry



Ivezić et al. (2012)

Photometric redshifts

- Less accurate (than spectroscopy)
- Degeneracy
- Much cheaper
- Template fitting (physical) & machine learning (empirical)

Machine learning framework:

- Features: a handful of galaxy colors ($\approx$ 3 to 30)
- Label: redshift
- Regression problem

Target model

Training set — With spec-z

Parameters

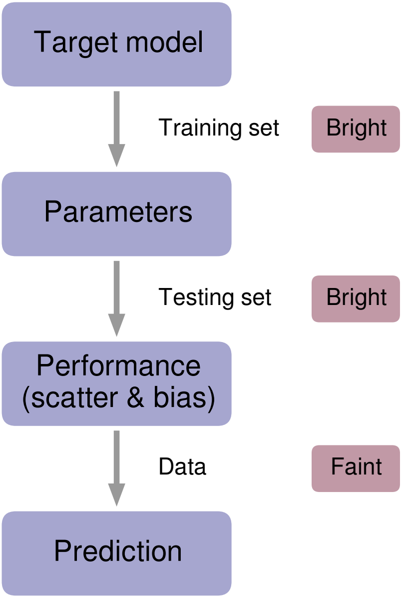
Testing set — With spec-z

Performance (scatter & bias)

Data — Without spec-z

Prediction

The training set
is not representative
(not enough)



The training set
is not representative
(not enough)

Let's design some tests

Target model

↓ Training set — Set A - SDSS DR12 spec

Parameters

↓ Testing set — Set B - SDSS DR12 spec / Set C - Intermediate case / Set D - SDSS DR12 photo

Performance (scatter & bias)

↓ Data

Prediction

Two catalogues
each with 4 subsets

Let’s design some tests

Target model

↓ Training set — Set A - SDSS DR12 spec

Parameters

↓ Testing set — Set B - SDSS DR12 spec; Set C - Intermediate case; Set D - SDSS DR12 photo

Performance (scatter & bias)

↓ Data

Prediction

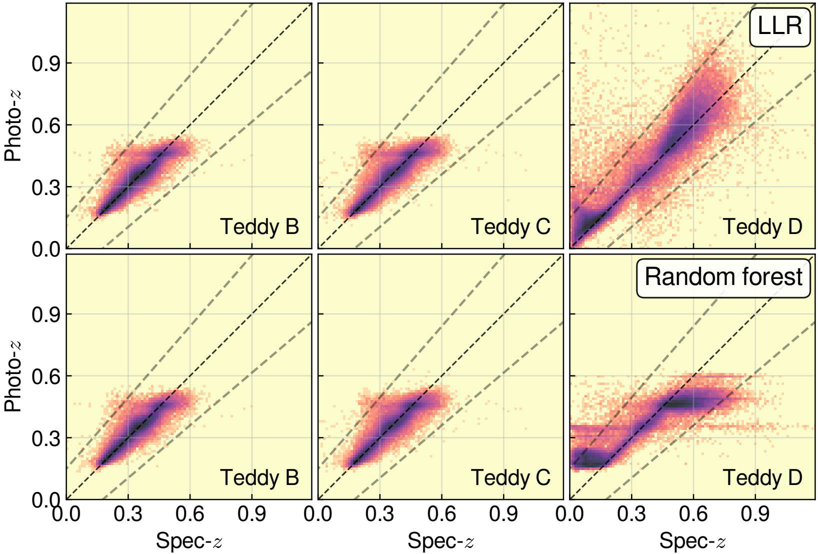
Two catalogues
each with 4 subsets

“Teddy”
Feature coverage

“Happy”
Feature uncertainty

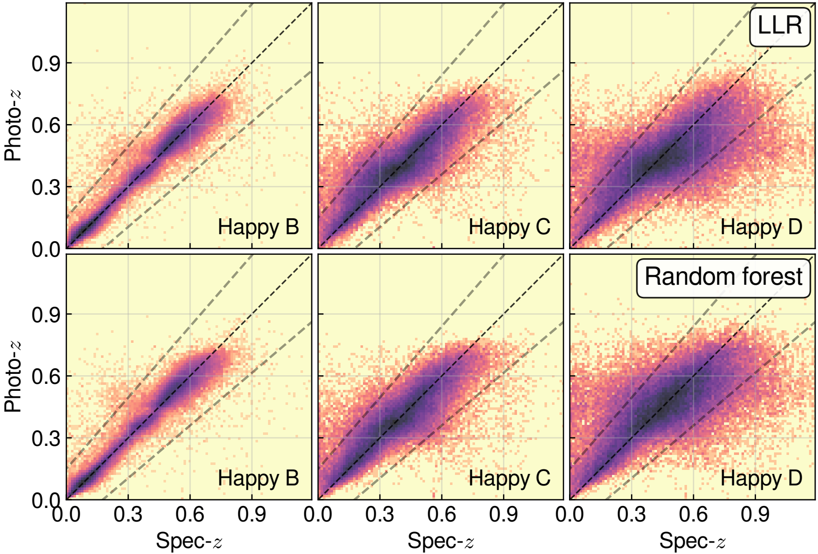
Beck, Lin, et al. (2017)

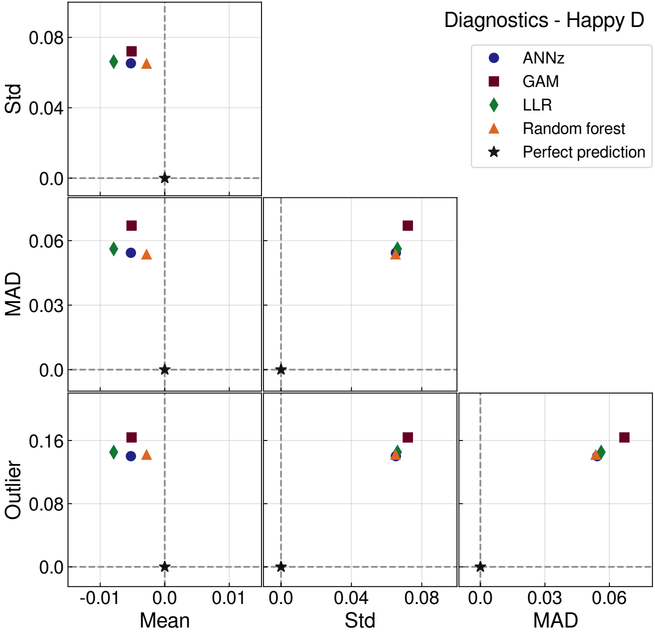
Impacts from non-representative coverage



Beck, Lin, et al. (2017)

Impacts from non-representative errors





Quantitative results

$$b \equiv \frac{z_{\text{photo}} - z_{\text{spec}}}{1 + z_{\text{spec}}}$$

MAD $= \text{median}(|b|)$
Outliers: $|b| > 0.15$

Beck, Lin, et al. (2017)

Solutions

Need for more representative training sets
$\Rightarrow$ Spectroscopic follow-ups

However, faint galaxies are expensive to follow up
$\Rightarrow$ The representativeness will never be perfect
$\Rightarrow$ Follow-ups to be optimized

Solutions

Need for more representative training sets

⇒ Spectroscopic follow-ups

However, faint galaxies are expensive to follow up

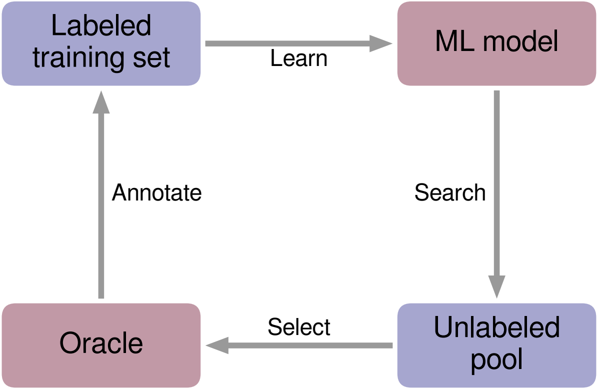
⇒ The representativeness will never be perfect

⇒ Follow-ups to be optimized

Active learning

(Let data tell us what to query)

Pool-based active learning



Active learning for photo-z

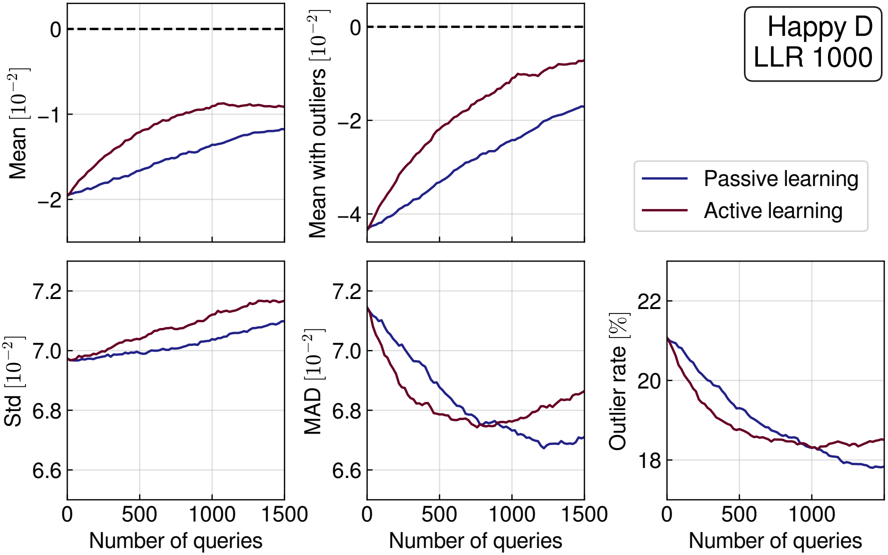
Initial training set	10000 samples from Set A (spec-like)
Unlabeled pool	10000 samples from Set D (photo-like)
Testing set	Another 30000 samples from Set D
Query strategy	Expected model change

Steps:

- Make a temporary reduction of the training set
- Compare prediction results between the full & the reduced cases
- Query the sample with the largest change

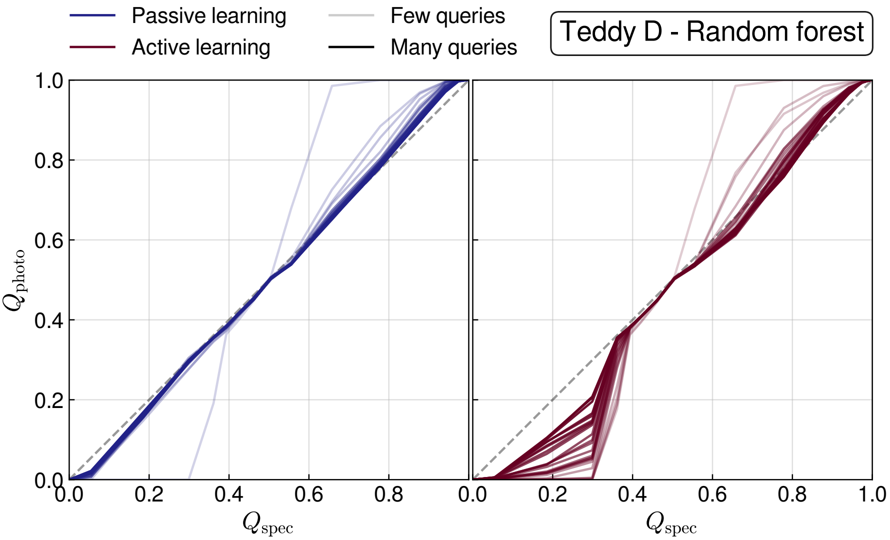
In prep.

Diagnostic results



In prep.

Cumulative distributions



Summary

- Naive validation of ML algorithms underestimates the photo-z scatter & bias.
- Color coverage & magnitude errors are major causes.
- Active learning would help debiasing the training set.

COIN

Cosmostatistics Initiative

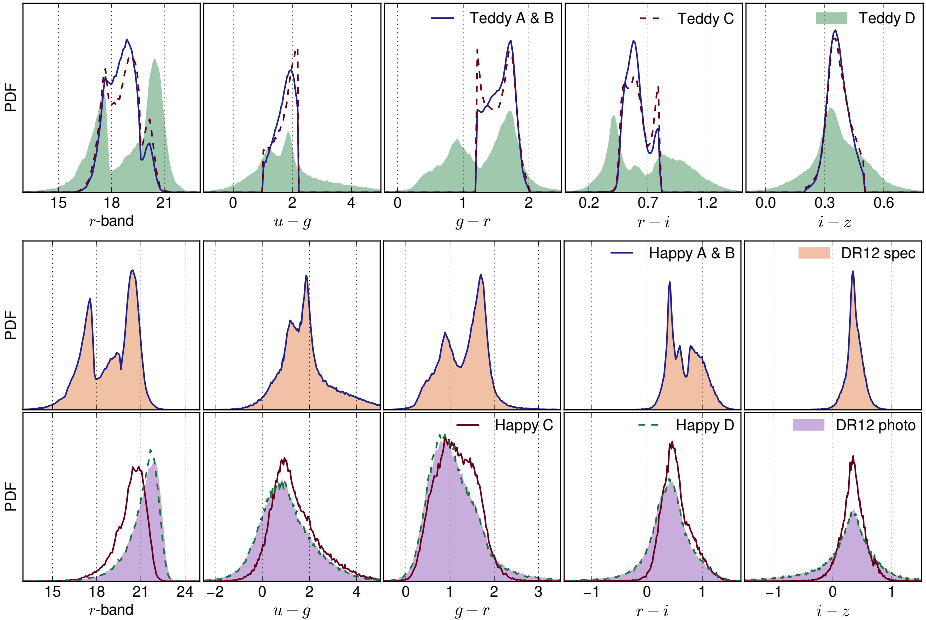
Catalogues available on GitHub

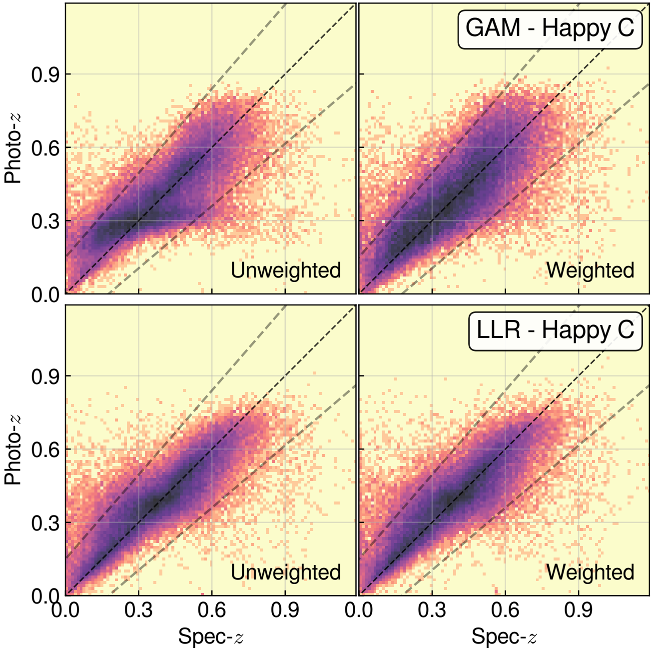
https://github.com/COINtoolbox/photoz_catalogues

Backup slides

Beck, Lin et al. (2017)

Color distributions

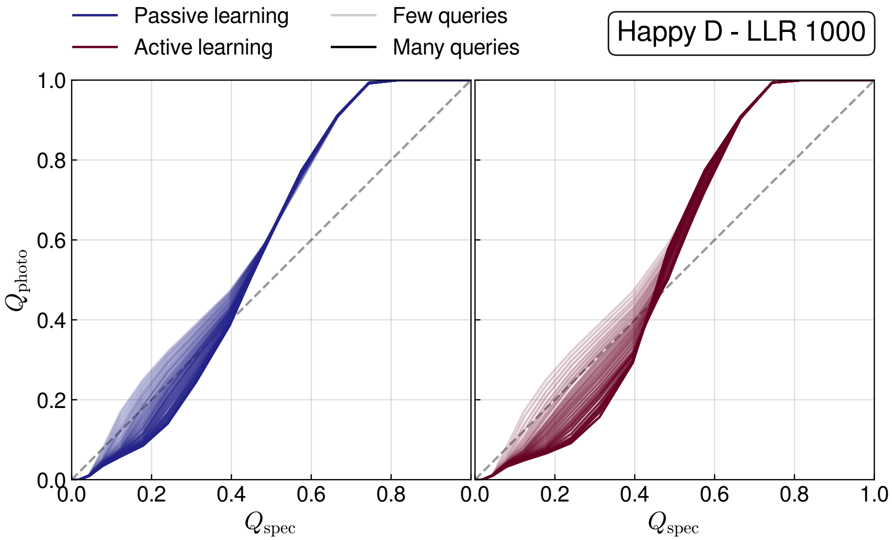




Reweighting does not cure

In prep.

Cumulative distributions



Method	Set	Mean	Std	MAD	Outliers
ANNz	B	0.04	2.87	1.49	0.99
	C	0.16	5.41	3.60	5.59
	D	-0.52	6.53	5.44	14.01
GAM	B	0.09	3.50	1.95	1.36
	C	0.86	6.34	4.84	7.37
	D	-0.51	7.21	6.70	16.38
LLR	B	0.13	2.81	1.39	1.11
	C	0.52	5.45	3.59	6.07
	D	-0.79	6.62	5.62	14.52
Random forest	B	0.05	2.82	1.41	1.02
	C	0.34	5.39	3.51	5.58
	D	-0.28	6.51	5.36	14.2

Diagnostics

Units: 10^{-2}

Extended spectroscopic samples

Survey	References	Number of matches
2dF	Colless et al. (2001, 2003)	770
6dF	Jones et al. (2004, 2009)	765
DEEP2	Davis et al. (2003); Newman et al. (2013)	7456
GAMA	Driver et al. (2011); Baldry et al. (2014)	53373
PRIMUS	Coil et al. (2011); Cool et al. (2013)	32459
VIPERS	Garilli et al. (2014); Guzzo et al. (2014)	18967
VVDS	Le Fèvre et al. (2004); Garilli et al. (2008)	8381
WiggleZ	Drinkwater et al. (2010); Parkinson et al. (2012)	43874
zCOSMOS	Lilly et al. (2007, 2009)	2789
Total		168834